Deriving LFG resource from Treebanks: Automatic Annotation of Treebank-Derived PSGs

April 13, 2000

Treebanks which encode higher-level functional structure, in addition to pure phrase structure information, are required as training resources for probabilistic unification grammars and data-driven parsing approaches, e.g. (Bod and Kaplan 1998). Manual construction of such treebanks is very labour and cost intensive. As an alternative, one could envisage the construction of new or the scaling-up of existing unification grammars which could then be used to analyze corpora. However, both scaling-up and grammar development is labour and cost intensive. What is more, even if a large-coverage unification grammar is available, typically, for each sentence it would come up with hundreds or thousands of candidate analyses from which a highly trained expert has to select. Although proposals have been made for filtering and ranking parsing ambiguities (e.g. (Frank et al. 1998)), to date none is guaranteed to uniquely determine the best analysis. In order not to compromise the quality of the corpus under construction a linguistic expert is required to find the best among a large number of candidate analyses.

We present a method for automatically annotating treebank resources with functional structures. The method defines systematic patterns of correspondence between partial PS configurations and functional structures and is applied to PS rules extracted from treebanks.

The basic idea involves taking an existing treebank, reading off the CF-PSG following (Charniak 1996), annotating it with f-structure annotations (providing macros for the lexical entries) and "reparsing" the treebanked trees simply following the original c-structure annotations. During this reparsing process, the f-structure annotations are resolved, and an f-structure is produced. The process is deterministic if the annotations are, and to a large extent costly manual inspection of candidate analyses is avoided.

The work of annotating the set of CFG rules derived from the treebank would be labour-intensive if carried out manually. Much recent work in LFG has shown that the c-structure f-structure correspondence for a configurational, generally endocentric language such as English, is largely predictable from a small set of mapping principles. In general, the correspondence between c-structure and f-structure follows linguistically determined principles which are partly universal, and partly language specific (Bresnan 1999), (Dalrymple 1999).

In the light of this, an obvious question to pursue is the extent to which principles can be stated for the automatic f-structure annotation of CFG rules derived from treebank representations. This can be expected to cast light on the soundness, accuracy and appropriacy of the linguists' generalisations: that is, the automatic procedure (applied to a large treebank derived ruleset) can serve as a testbed for the linguistic principles.

In this paper, we show how f-structure annotation of treebanks and grammar rules can (to a large extent) be automated. This result substantially extends the research in (van Genabith et al. 1999c, van Genabith et al. 1999a, van Genabith et al. 1999b).

In the paper we present the method and discuss its application to a grammar (of around 500 rules) derived from a fragment of the AP treebank. We give some results on precision and recall. We then go on to discuss several linguistic issues concerning treebank representations.

Bibliography

Bod, R., and R. Kaplan. 1998. A probabilistic corpus-driven model for lexical-functional analysis. In *Proceedings of COLING/ACL'98*, 145–151.

Bresnan, J. 1999. Lexical-functional syntax. Manuscript, Stanford University.

Charniak, E. 1996. Tree-bank grammars. In *AAAI-96. Proceedings of the Thirteenth National Conference on Artificial Intelligence*, 1031–1036. MIT Press.

Dalrymple, M. 1999. Lexical-functional grammar. Manuscript, Xerox PARC.

Frank, A., T. King, J. Kuhn, and J. Maxwell-III. 1998. Optimality theory style constraint ranking in large-scale lfg grammars. In M. Butt and T. King (Eds.), *Proceedings of the LFG98 Conference*, University of Queensland, Brisbane, CSLI Online Publications, Stanford, CA. http://www-csli.stanford.edu/publications/.

van Genabith, J., L. Sadler, and A. Way. 1999a. Data-driven compilation of lfg semantic forms. In *EACL'99 Workshop on Linguistically Interpreted Corpora (LINC-99), Bergen, Norway, June 12th*, 69–76.

van Genabith, J., L. Sadler, and A. Way. 1999b. Structure preserving cf-psg compaction, lfg and treebanks. In *Proceedings ATALA Workshop - Treebanks*, Journees ATALA, Corpus annotes pour la syntaxe, Universite Paris 7, France, 18-19 Juin 1999, 107–114.

van Genabith, J., A. Way, and L. Sadler. 1999c. Semi-automatic generation of f-structures from tree banks. In M. Butt and T. King (Eds.), *Proceedings of the LFG99 Conference*, Manchester University, 19-21 July, CSLI Online Publications, Stanford, CA. http://www-csli.stanford.edu/publications/.